LOHAI: Providing a baseline for KOS based automatic indexing

Kai Eckert

Mannheim University Library, Germany
eckert@bib.uni-mannheim.de

First workshop on
Semantic Digital Archives (SDA 2011),
Sep 29th 2011, Berlin

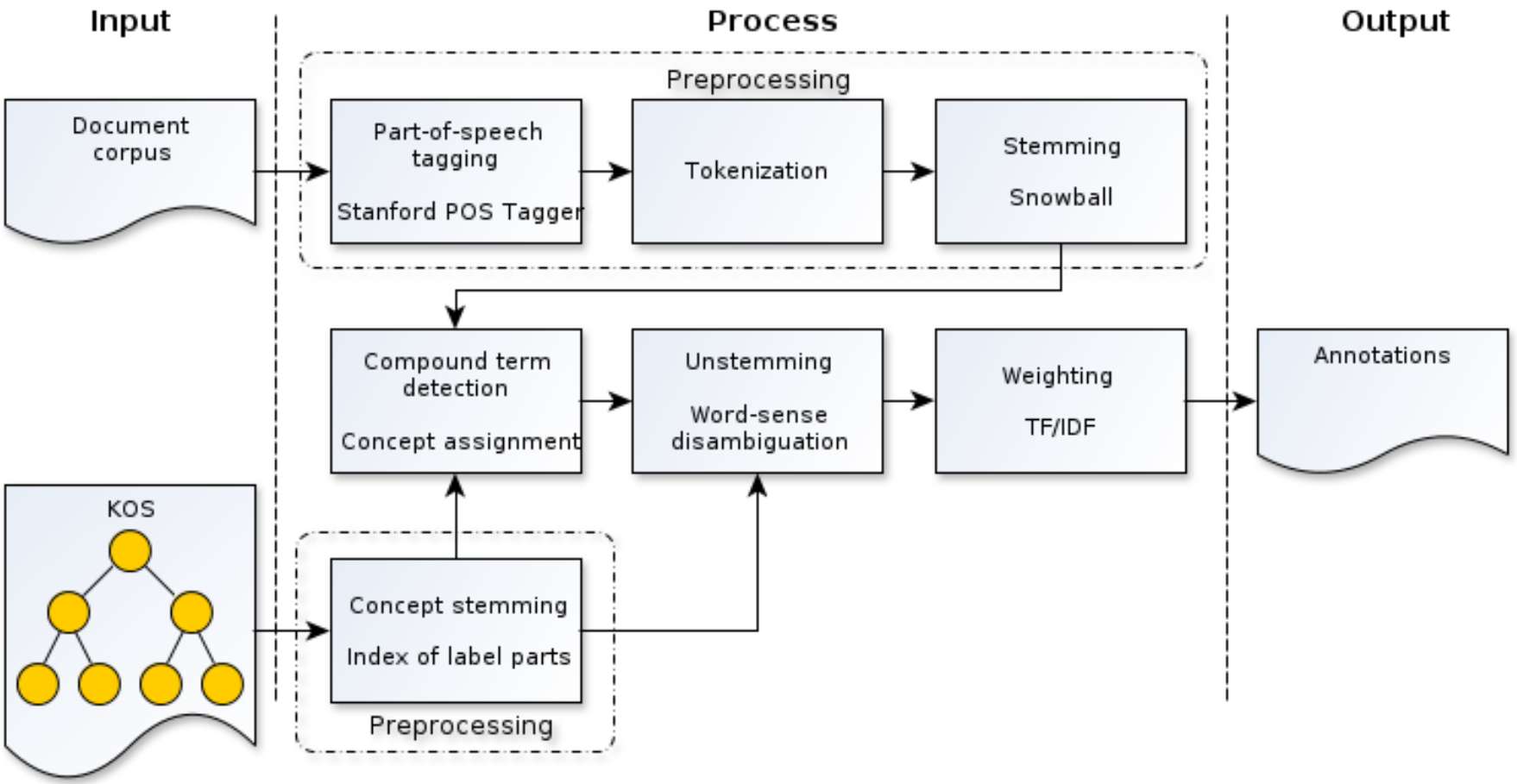
Motivation

- General KOS based indexer for various (even serious) purposes:
  - Document exploration
  - Thesaurus examination
  - Automatic Indexing
  - ...
- No free and open source implementation was available.
- LOHAI: **Lo**w **Ha**nging Fruits Automatic **I**ndexer

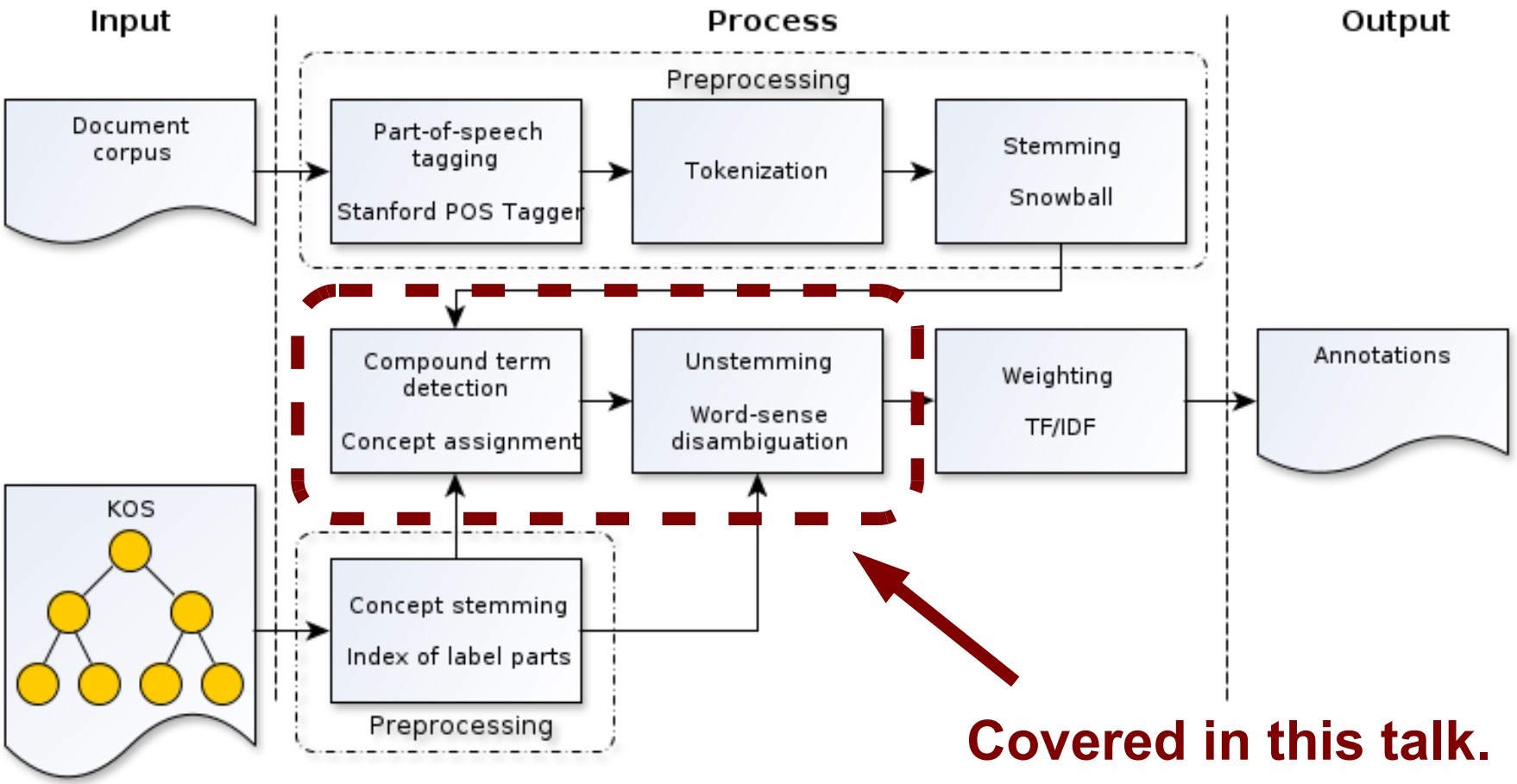
Design Principles

- **Simplicity over quality**
  - Easy to **use**
  - Easy to **understand**
  - Easy to **improve**
- **Knowledge-poor** and **without any training**
  - Must not rely on any **additional sources**
  - **No training step**, to be usable in a setting where no preindexed documents are available.

Indexing Pipeline



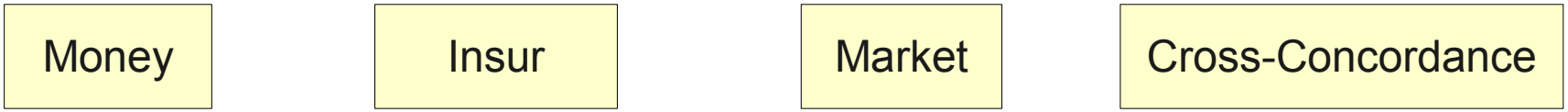
Indexing Pipeline



Required Disambiguation

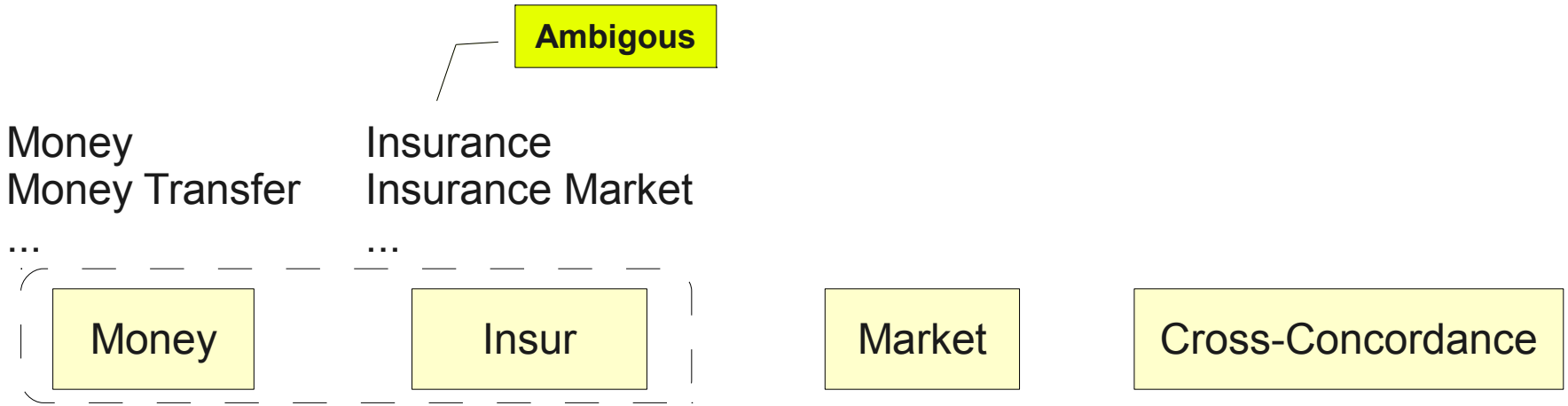
- **Compound terms:**
  - “insur” >> “insurance”, “insurance market”
- **Overstemming:**
  - „nation“ >> “nationalism”, “nationality”, “nation”
- **Homonyms:**
  - “bank”: *the financial institution*
  - “bank”: *a raised portion of seabed or sloping ground along the edge of a stream, river, or lake*

Compound Term Detection

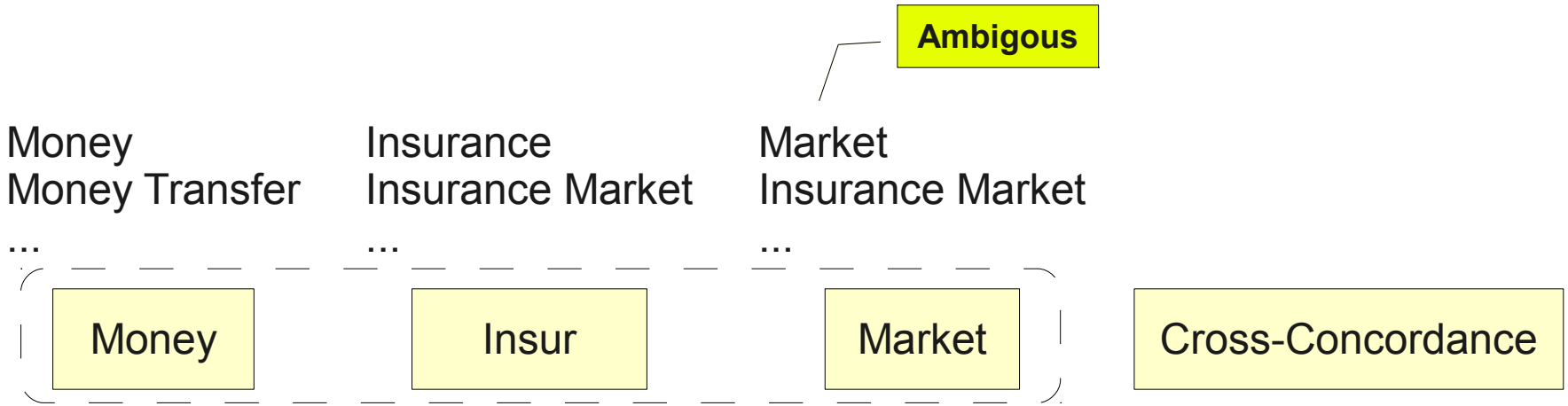


Compound Term Detection

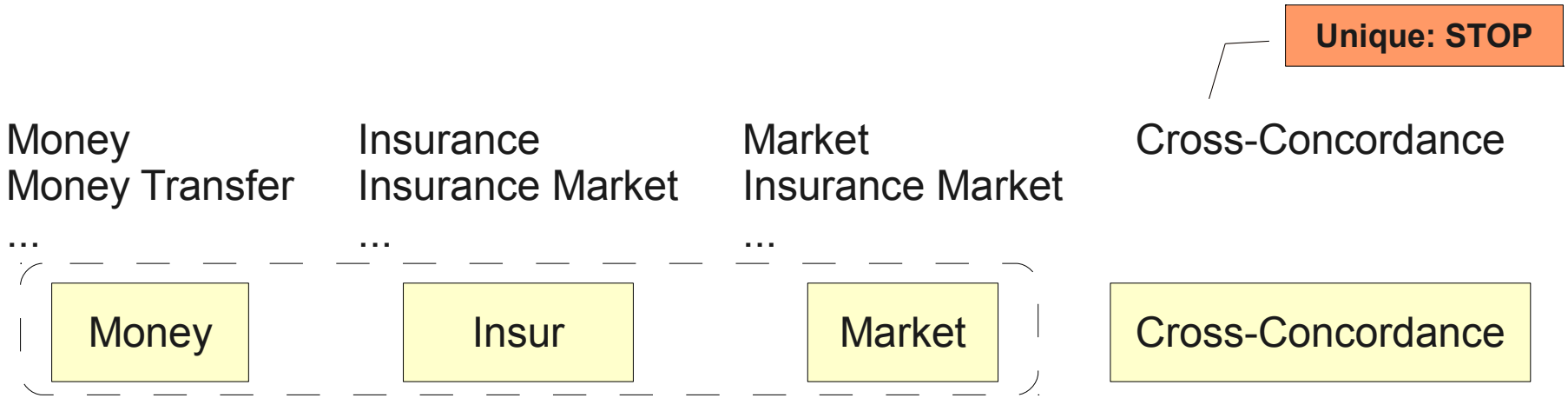
Compound Term Detection



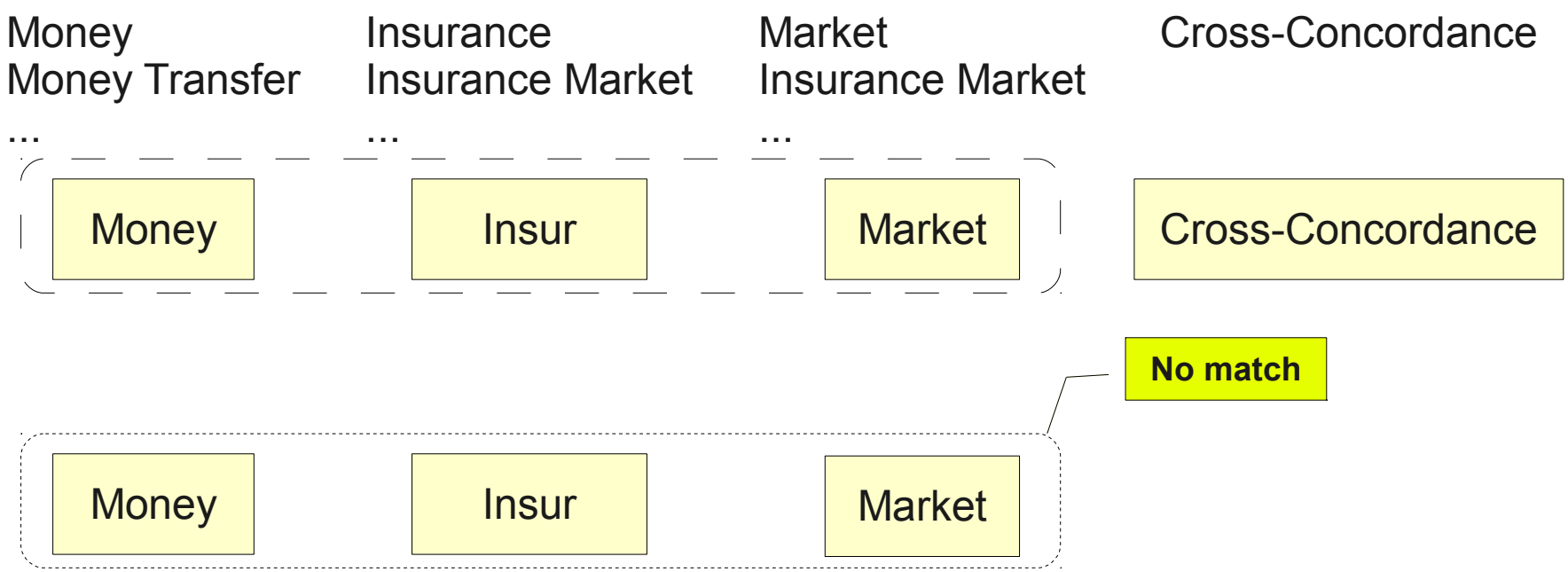
Compound Term Detection



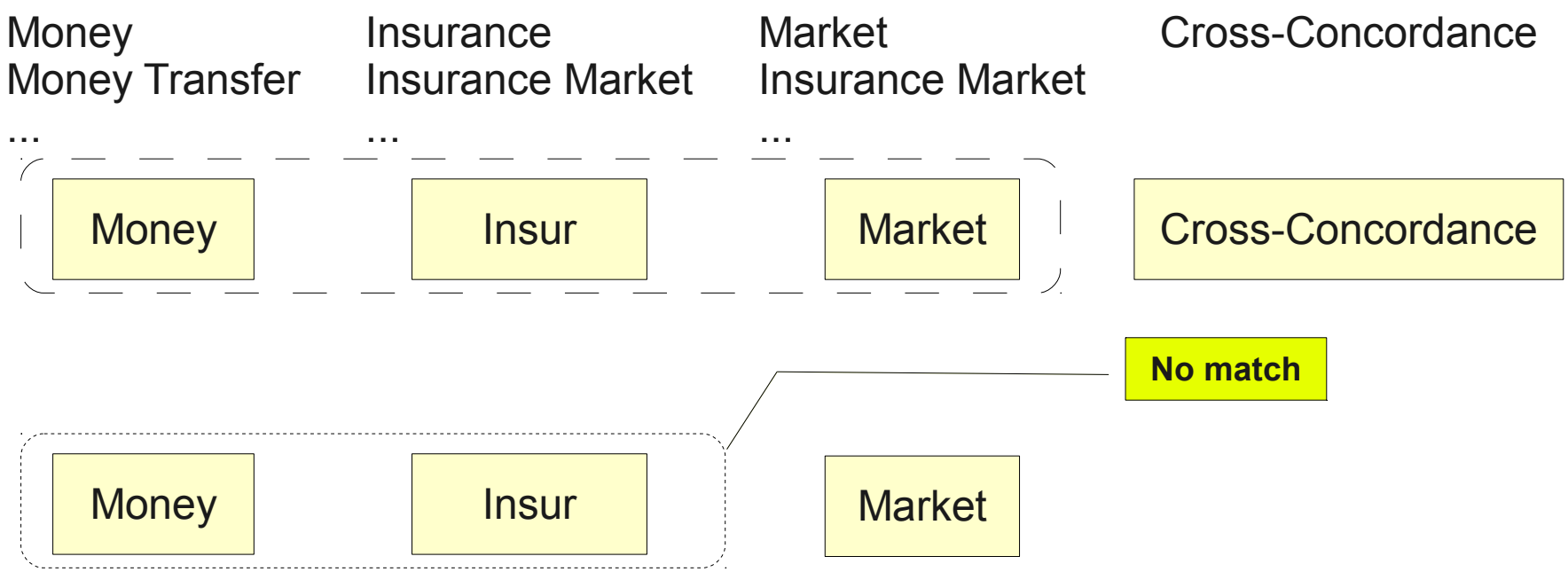
Compound Term Detection



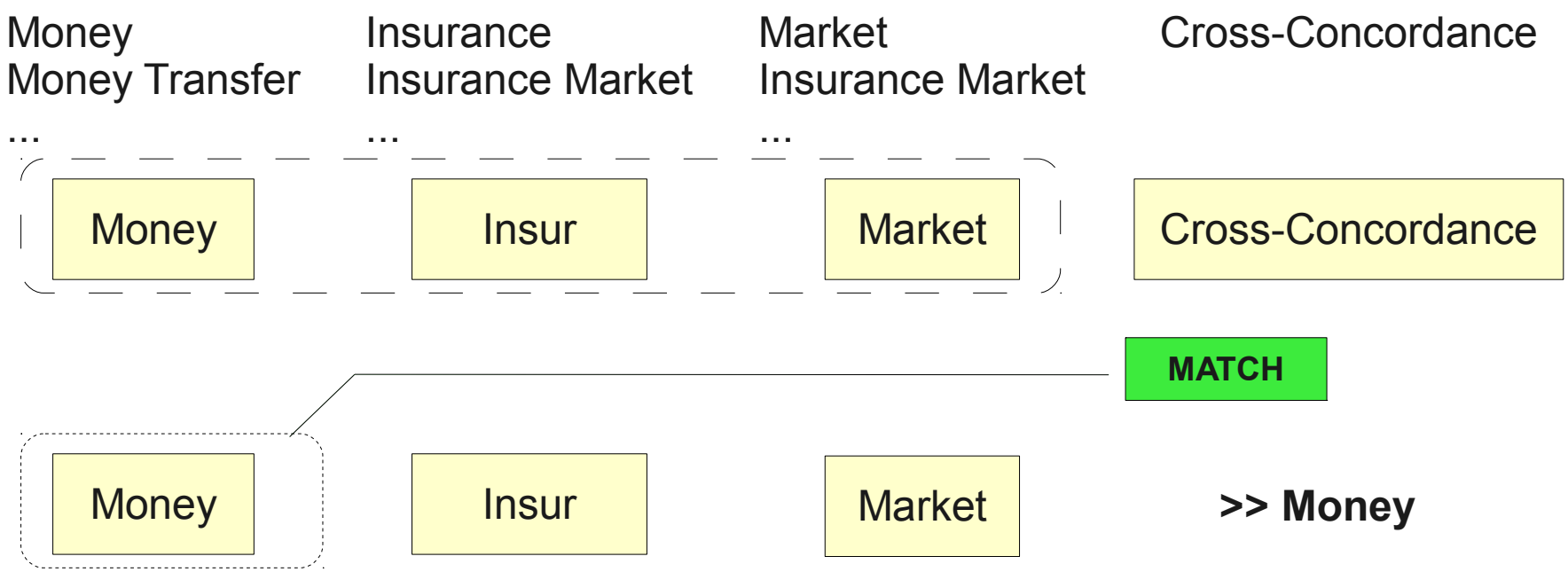
Compound Term Detection



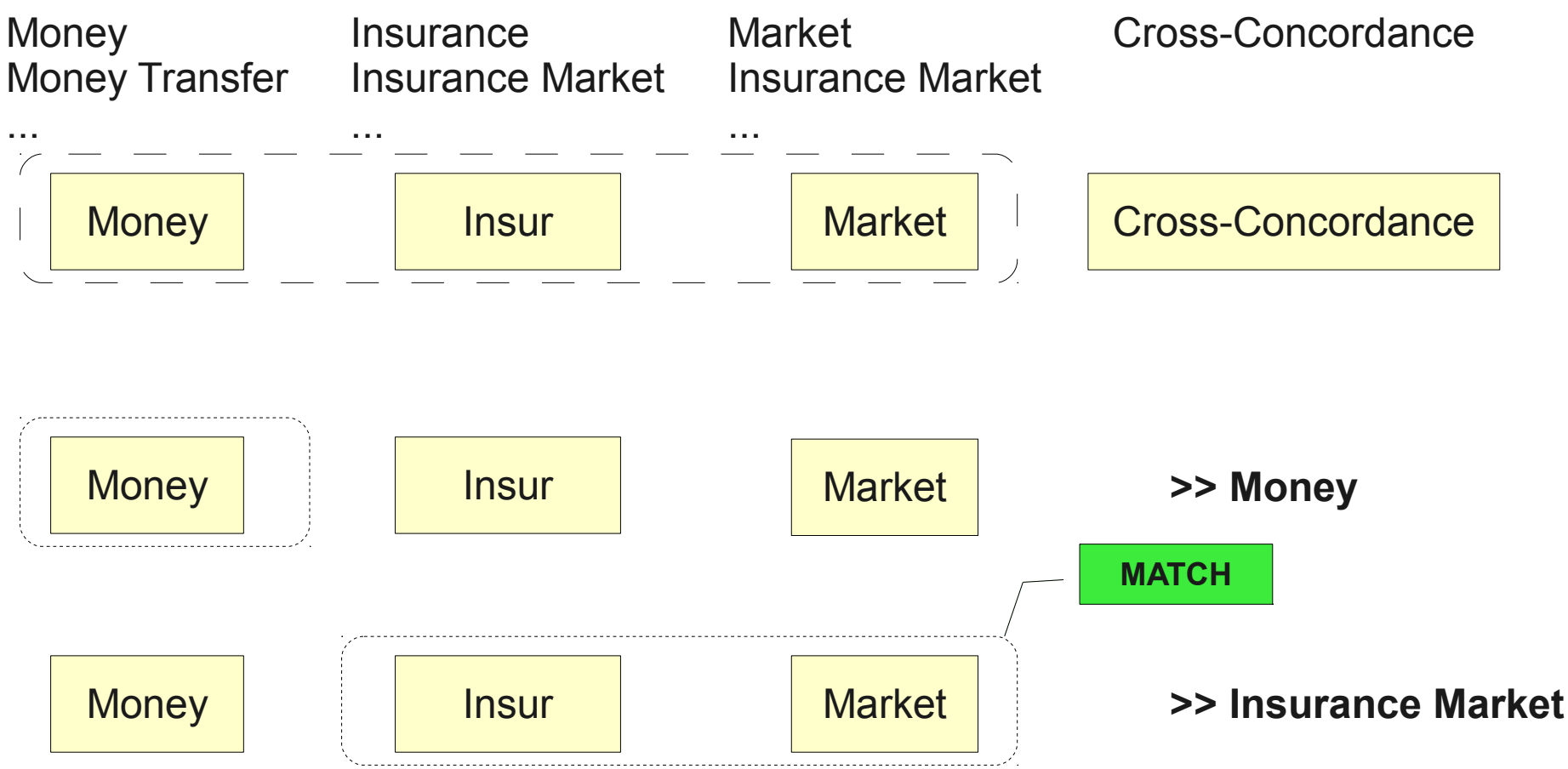
Compound Term Detection



Compound Term Detection



Compound Term Detection



Unstemming

- Overstemming:
  - „nation“ >> “nationalism”, “nationality”, “nation”
- Compare **unstemmed terms**.
- If they match:
  - Assign them.
- If not:
  - Continue with **Word Sense Disambiguation**.

Word Sense Disambiguation

Yarowsky's assumptions:

- **One sense per collocation**
  - Collocated terms are unique for each possible sense of a given term.
- **One sense per discourse**
  - Only one sense for a given word is used throughout a whole document.

Jaccard Measure

- Term Environments:
  - Document: **100 words** before and after the term occurrence form set ***W***.
  - KOS: All labels of the concept, its direct **children**, **siblings** and **parents** form set ***C***.
- Jaccard Measure:

$$Jaccard(W, C) = \frac{|W \cup C|}{|W \cap C|}$$

- Assign one sense per document, based on the Jaccard value of **all** occurrences.

Example Result

Hardin, Russell: *Contractarianism: Wistful Thinking*

The contract metaphor in political and moral theory is misguided. It is a poor metaphor both descriptively and normatively, but here I address its normative problems. Normatively, contractarianism is supposed to give justifications for political institutions and for moral rules, just as contracting in the law is supposed to give justification for claims of obligation based on consent or agreement. This metaphorical association fails for several reasons. First, actual contracts generally govern prisoner's dilemma, or exchange, relations; the so-called social contract governs these and more diverse interactions as well. Second, agreement, which is the moral basis of contractarianism, is not right-making per se. Third, a contract in law gives information on what are the interests of the parties; a hypothetical social contract requires such knowledge, it does not reveal it. Hence, much of contemporary contractarian theory is perversely rationalist at its base because it requires prior, rational derivation of interests or other values. Finally, contractarian moral theory has the further disadvantage that, unlike contract in the law, its agreements cannot be connected to relevant motivations to abide by them.

Constitutional Political Economy, 1 (2) 1990: 35-52

Example Result

Manual assignment

- Constitutional economics
- Influence of government
- Ethics
- Theory

LOHAI

- Contract Law (1.21)
- Contract (0.76)
- Social contract (0.64)
- Law (0.51)
- Politics (0.37)
- Prisoner's dilemma (0.34)
- Theory (0.32)

- Rationalism (0.24)
- Association (0.23)
- Exchange (0.20)
- Knowledge (0.19)
- Government (0.16)
- Information (0.12)

Example Result

The contract metaphor in political and moral theory is misguided. It is a poor metaphor both descriptively and normatively, but here I address its normative problems. Normatively, contractarianism is supposed to give justifications for political institutions and for moral rules, just as contracting in the law is supposed to give justification for claims of obligation based on consent or agreement. This metaphorical association fails for several reasons. First, actual contracts generally govern prisoner's dilemma, or exchange, relations; the so-called social contract governs these and more diverse interactions as well. Second, agreement, which is the moral basis of contractarianism, is not right-making per se. Third, a contract in law gives information on what are the interests of the parties; a hypothetical social contract requires such knowledge, it does not reveal it. Hence, much of contemporary contractarian theory is perversely rationalist at its base because it requires prior, rational derivation of interests or other values. Finally, contractarian moral theory has the further disadvantage that, unlike contract in the law, its agreements cannot be connected to relevant motivations to abide by them.

Conclusion

- Reasonable results without „Black Box“ Effect.
- Directly usable for new KOS and document sets.
- No training step needed.
- Low hanging fruits, but a good baseline.
- ~ 500 LoC in (quite verbose) Java.
- Easy to adapt and to improve.
- Free and open source:
  - https://github.com/kaiec/LOHAI

Thank you.

Questions/Ideas?

Kai Eckert
Mannheim University Library, Germany

eckert@bib.uni-mannheim.de